TOPMed Whole Genome Sequencing Project

October 9, 2016

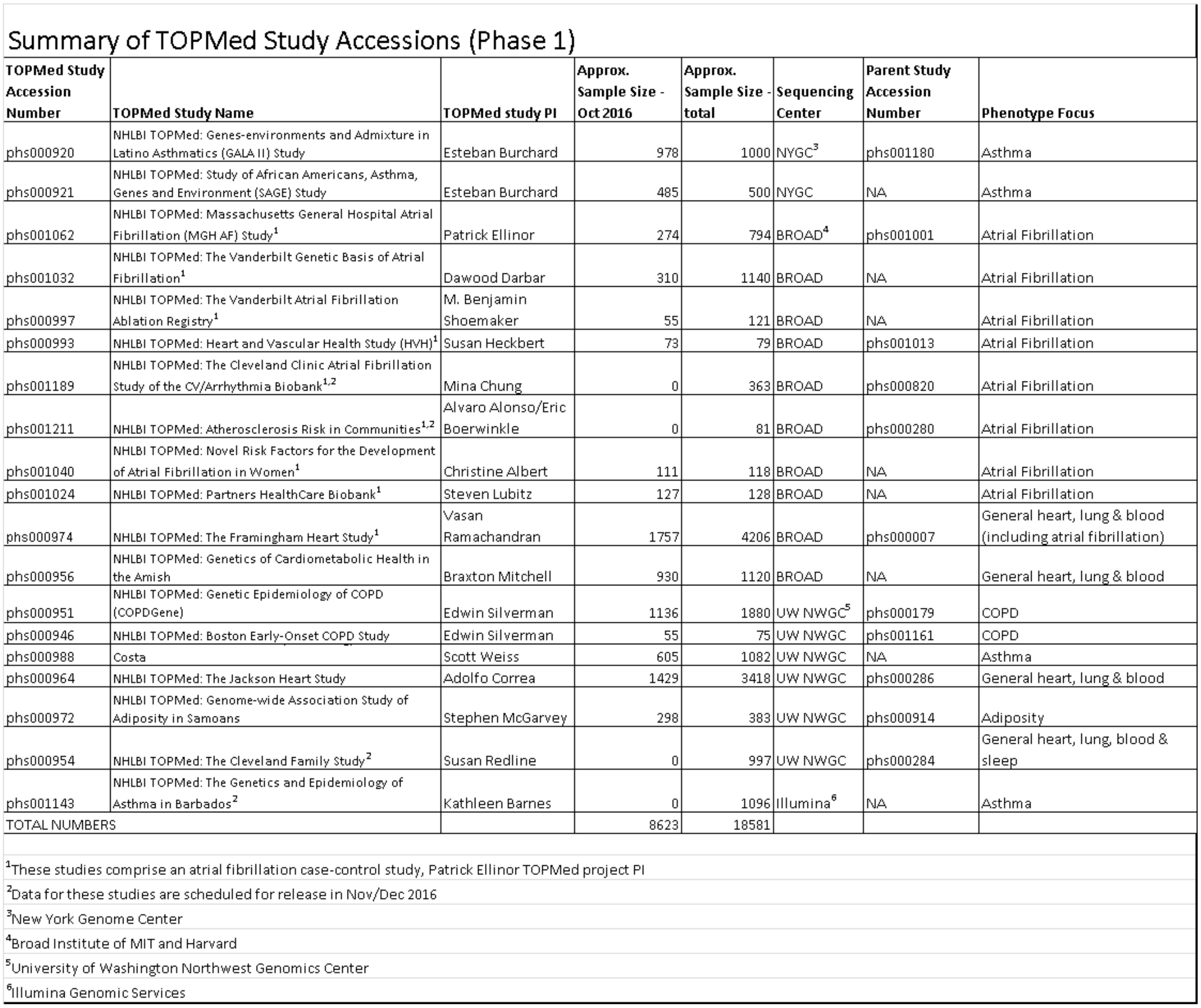
Introduction

Overview

Trans-Omics for Precision Medicine (TOPMed), sponsored by the National Heart, Lung and Blood Institute (NHLBI), generates scientific resources to enhance our understanding of fundamental biological processes that underlie heart, lung, blood and sleep disorders (HLBS). It is part of the broader Precision Medicine Initiative, which aims to provide disease treatments that are tailored to an individual's unique genes and environment. TOPMed contributes to this initiative by integrating whole-genome sequencing (WGS) and other –omics data (e.g., metabolic profiles, protein and RNA expression patterns) with molecular, behavioral, imaging, environmental, and clinical data. In doing so, the TOPMed program seeks to uncover factors that increase or decrease the risk of disease, identify subtypes of disease, and develop more targeted and personalized treatments.

Currently, TOPMed includes 26 different studies with ~72,000 samples undergoing whole genome sequencing. The studies encompass several experimental designs (e.g. cohort, case-control, family) and many different clinical trait areas (e.g. asthma, COPD, atrial fibrillation, atherosclerosis, sleep). See study descriptions under the "Groups" tab on the TOPMed web site (www.nhlbiwgs.org).

TOPMed WGS data will be released in multiple waves. The first release, in October 2016, will include ~8,600 samples in 15 separate dbGaP accessions, followed by four additional accessions in Nov/Dec 2016. These accessions are summarized in the Table below. Some TOPMed studies have previously released genotypic and phenotypic data on dbGaP in "parent" accessions (see Table). For those studies, the TOPMed WGS accession contains only WGS-derived data and, therefore, genotype-phenotype analysis requires access to data from both parent and TOPMed WGS accessions. For the studies in Table 1 without a specific parent accession number, the TOPMed WGS accession contains both genotype and phenotype data.



Summary of TOPMed Study Accessions (Phase 1)

TOPMed Study Accession Number	TOPMed Study Name	TOPMed study PI	Approx. Sample Size - Oct 2016	Approx. Sample Size - total	Sequencing Center	Parent Study Accession Number	Phenotype Focus
phs000920	NHLBI TOPMed: Genes-environments and Admixture in Latino Asthmatics (GALA II) Study	Esteban Burchard	978	1000	NYGC[3]	phs001180	Asthma
phs000921	NHLBI TOPMed: Study of African Americans, Asthma, Genes and Environment (SAGE) Study	Esteban Burchard	485	500	NYGC	NA	Asthma
phs001062	NHLBI TOPMed: Massachusetts General Hospital Atrial Fibrillation (MGH AF) Study[1]	Patrick Ellinor	274	794	BROAD[4]	phs001001	Atrial Fibrillation
phs001032	NHLBI TOPMed: The Vanderbilt Genetic Basis of Atrial Fibrillation[1]	Dawood Darbar	310	1140	BROAD	NA	Atrial Fibrillation
phs000997	NHLBI TOPMed: The Vanderbilt Atrial Fibrillation Ablation Registry[1]	M. Benjamin Shoemaker	55	121	BROAD	NA	Atrial Fibrillation
phs000993	NHLBI TOPMed: Heart and Vascular Health Study (HVH)[1]	Susan Heckbert	73	79	BROAD	phs001013	Atrial Fibrillation
phs001189	NHLBI TOPMed: The Cleveland Clinic Atrial Fibrillation Study of the CV/Arrhythmia Biobank[1,2]	Mina Chung	0	363	BROAD	phs000820	Atrial Fibrillation
phs001211	NHLBI TOPMed: Atherosclerosis Risk in Communities[1,2]	Alvaro Alonso/Eric Boerwinkle	0	81	BROAD	phs000280	Atrial Fibrillation
phs001040	NHLBI TOPMed: Novel Risk Factors for the Development of Atrial Fibrillation in Women[1]	Christine Albert	111	118	BROAD	NA	Atrial Fibrillation
phs001024	NHLBI TOPMed: Partners HealthCare Biobank[1]	Steven Lubitz	127	128	BROAD	NA	Atrial Fibrillation
phs000974	NHLBI TOPMed: The Framingham Heart Study[1]	Vasan Ramachandran	1757	4206	BROAD	phs000007	General heart, lung & blood (including atrial fibrillation)
phs000956	NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish	Braxton Mitchell	930	1120	BROAD	NA	General heart, lung & blood
phs000951	NHLBI TOPMed: Genetic Epidemiology of COPD (COPDGene)	Edwin Silverman	1136	1880	UW NWGC[5]	phs000179	COPD
phs000946	NHLBI TOPMed: Boston Early-Onset COPD Study	Edwin Silverman	55	75	UW NWGC	phs001161	COPD
phs000988	Costa	Scott Weiss	605	1082	UW NWGC	NA	Asthma
phs000964	NHLBI TOPMed: The Jackson Heart Study	Adolfo Correa	1429	3418	UW NWGC	phs000286	General heart, lung & blood
phs000972	NHLBI TOPMed: Genome-wide Association Study of Adiposity in Samoans	Stephen McGarvey	298	383	UW NWGC	phs000914	Adiposity
phs000954	NHLBI TOPMed: The Cleveland Family Study[2]	Susan Redline	0	997	UW NWGC	phs000284	General heart, lung, blood & sleep
phs001143	NHLBI TOPMed: The Genetics and Epidemiology of Asthma in Barbados[2]	Kathleen Barnes	0	1096	Illumina[6]	NA	Asthma
TOTAL NUMBERS			8623	18581			

[1] These studies comprise an atrial fibrillation case-control study, Patrick Ellinor TOPMed project PI

[2] Data for these studies are scheduled for release in Nov/Dec 2016

[3] New York Genome Center

[4] Broad Institute of MIT and Harvard

[5] University of Washington Northwest Genomics Center

[6] Illumina Genomic Services

The following sections of this document describe methods of data acquisition, processing and quality control (QC) for TOPMed WGS data contained in the 2016 releases. Briefly, ~30X whole genome sequencing was performed at several different Sequencing Centers (named in the Table). All samples for a given study were sequenced at the same center, except for a small number of control samples described below. The reads were aligned to human genome build GRCh37 at each center using similar, but not identical, processing pipelines. The resulting BAM files were transferred from all centers to the TOPMed Informatics Research Center (IRC), where they were re-aligned to build GRCh37, using a common pipeline to produce a set of 'harmonized' BAM files. Both the Sequencing Center-specific BAM and the harmonized BAM files were deposited in the NCBI Sequence Read Archive (SRA), where they were converted to '.sra' file format. Both center-specific and IRC-harmonized .sra files are available to users with approved access to a given study. The IRC performed joint genotype calling on all samples in the October 2016 releases (along with additional samples to be released later). The resulting VCF files were split by study and consent group for distribution to approved dbGaP users, but can be reassembled easily for cross-study, pooled analysis because the files for all studies contain the same variant sites. Quality control was performed at each stage of the process by the Sequencing Centers, the IRC and the TOPMed Data Coordinating Center (DCC). Only samples and variants that passed QC are included in the genotype call sets distributed with the 2016 releases.

Sequence/genotype data files provided in the 2016 dbGaP releases include the following:

1. Aligned read data for each sample in '.sra' format (which is readily convertible to BAM format). Each sample has two .sra files: one from the Sequencing Center and the other from the IRC
2. Genotype call sets (one per chromosome) in '.vcf' format

TOPMed DNA sample/sequencing-instance identifiers

Each DNA sample processed by TOPMed was given a unique identifier as "NWD" followed by six digits (e.g. NWD123456). These identifiers are unique across all TOPMed studies. Each NWD identifier is associated with a single study subject identifier used in other dbGaP files (such as phenotypes, pedigrees and consent files). A given subject identifier may link to multiple NWD identifiers when duplicate samples are taken from the same individual. Study investigators assigned NWD IDs to subjects, and their biorepositories assigned DNA samples/ NWD IDs to specific bar-coded wells/tubes supplied by their Sequencing Center, and recorded those assignments in a sample manifest, along with other metadata (e.g. sex, DNA extraction method). At each Sequencing Center, the NWD ID was propagated through all phases of the pipeline and is the primary identifier in all results files. Each NWD ID resulted in a single sequencing instance (i.e. 'run' in SRA terminology).

Control Samples

One parent-offspring trio from the Framingham Heart Study (FHS) was sequenced at each of four Sequencing Centers (family ID 746, subject IDs 13823, 15960 and 20156). All four WGS runs for each subject are provided in the TOPMed FHS accession (phs000974). In addition, HapMap subjects NA12878 (CEU, Lot K6) and NA19238 (YRI, Lot E2) were sequenced at each of the Sequencing Centers in alternation, once approximately every 1000 study samples. The HapMap sequence data will be released publicly as a BioProject in Q4 2016 or Q1 2017.

One parent-offspring trio from the Framingham Heart Study (FHS) was sequenced at each of four Sequencing Centers (family ID 746, subject IDs 13823, 15960 and 20156). All four WGS runs for each subject are provided in the TOPMed FHS accession (phs000974). In addition, HapMap subjects NA12878 (CEU, Lot K6) and NA19238 (YRI, Lot E2) were sequenced at each of the Sequencing Centers in alternation, once approximately every 1000 study samples. The HapMap sequence data will be released publicly as a BioProject in Q4 2016 or Q1 2017.

The average pairwise non-reference genotype discordance rate among 69 pairs of duplicate sequenced samples is 5 x 10^{-5} on the set of variants included in this release. The genotype discordance rate is very sensitive to the stringency level used for variant site filtering. This low figure is evidence of the benefit of 30x whole genome sequencing and suggests that the current filtering threshold suitably balances sensitivity and specificity. It must be acknowledged that these 27 control samples were among 4,047 duplicate and related samples which provided a negative training set for the SVM classifier used for site level filtering (see Variant Filtering section).

To calculate non-reference discordance, the genotypes of each DNA sample are called independently from separate sets of sequence reads, often from different Sequencing Centers. The denominator for each pairwise comparison is the number of sites where at least one of the two samples has a non-reference genotype called (either het or hom-alt). The numerator is the number of sites where the two genotypes disagree.

Sequencing Center Methods

Broad Institute of MIT and Harvard

Stacey Gabriel

DNA Sample Handling and QC

DNA samples are informatically received into the Genomics Platform's Laboratory Information Management System via a scan of the tube barcodes using a Biosero flatbed scanner. This registers the samples and enables the linking of metadata based on well position. All samples are then weighed on a BioMicro Lab's XL20 to determine the volume of DNA present in sample tubes. Following this the samples are quantified in a process that uses PICO-green fluorescent dye. Once volumes and concentrations are determined, the samples are handed off to the Sample Retrieval and Storage Team for storage in a locked and monitored -20 walk-in freezer.

Library Construction

Samples undergo fragmentation by means of acoustic shearing using Covaris focused-ultrasonicator, targeting 385 bp fragments. Following fragmentation, additional size selection is performed using a SPRI cleanup. Library preparation is performed using a commercially available kit provided by KAPA Biosystems (product KK8202) with palindromic forked adapters with unique 8 base index sequences embedded within the adapter (purchased from IDT). Following sample preparation, libraries are quantified using quantitative PCR (kit purchased from KAPA biosystems) with probes specific to the ends of the adapters. This assay is automated using Agilent’s Bravo liquid handling platform. Based on qPCR quantification, libraries are normalized to 1.7 nM. Samples are then pooled into 24-plexes and the pools are once again qPCRed. Samples are then combined with HiSeq X Cluster Amp Mix 1,2 and 3 into single wells on a strip tube using the Hamilton Starlet Liquid Handling system.

Clustering and Sequencing

As described in the library construction process, 96 samples on a plate are processed together through library construction. A set of 24 barcodes is used to index the samples. Barcoding allows pooling of samples prior to loading on sequencers and mitigates lane-lane effects at a single sample level. The plate is broken up into 4 pools of 24-samples each. The four pools are taken as columns on the plate (e.g., columns 1-3; 4-6; 7-9; 10-12). From this format (and given the current yields of a HiSeqX) the 4 pools are spread over 3 flowcells (24 lanes). Cluster amplification of the templates was performed according to the manufacturer's protocol (Illumina) using the Illumina cBot. Flowcells were sequenced on Hi Seq X with sequencing software HiSeq Control Software (HCS) version 3.3.76, then analyzed using RTA2 (Real Time Analysis).

Read Processing

For TOPMED phase 1 data the following versions were used for aggregation, and alignment to hg19_decoy reference: picard (latest version available at the time of the analysis), GATK (3.1-144-g00f68a3) and BwaMem (0.7.7-r441).

Sequence Data QC

A sample is considered sequence complete when the mean coverage is >= 30x. Two QC metrics that are reviewed along with the coverage are the sample Fingerprint LOD score (score which estimates the probability that the data is from a given individual) and % contamination. At aggregation, we do an all-by-all comparison of the read group data and estimate the likelihood that each pair of read groups is from the same individual. If any pair has a LOD score < -20.00, the aggregation does not proceed and is investigated. FP LOD >= 3 is considered passing concordance with the sequence data (ideally we see LOD >10). A sample will have an LOD of 0 when the sample failed to have a passing fingerprint. Fluidigm fingerprint is repeated once if failed. Read groups with fingerprints < -3.00 are blacklisted from the aggregation. If the sample does not meet coverage, it will be topped off for additional coverage. If a large % of read groups are blacklisted, it will be investigated as a potential sample swap. In terms of contamination, a sample is considered passing if the contamination is less than 5%. In general, the bulk of the samples have less than 1% contamination.

Northwest Genomics Center

Deborah Nickerson

DNA Sample Handling and QC

The NWGC centralizes all receipt, tracking, and quality control/assurance of DNA samples in a Laboratory Information Management System. Samples are assigned unique barcode tracking numbers and have a detailed sample manifest (i.e., identification number/code, sex, DNA concentration, barcode, extraction method). Initial QC entails DNA quantification, sex typing, and molecular "fingerprinting" using a high frequency, cosmopolitan genotyping assay. This 'fingerprint' is used to identify potential sample handling errors and provides a unique genetic ID for each sample, which eliminates the possibility of sample assignment errors. In addition, we spot check ~8% of the samples per batch on an agarose gel to check for high molecular weight DNA, if DNA degradation is detected all samples are checked. Samples are failed if: (1) the total amount, concentration, or integrity of DNA is too low; (2) the fingerprint assay produces poor genotype data or (3) sex-typing is inconsistent with the sample manifest. Barcoded plates were shipped to Macrogen for library construction and sequencing.

Library Construction

Libraries were constructed with a minimum of 0.4ug gDNA and are prepared in Covaris 96 microTUBE plates and sheared through a Covaris LE220 focused ultrasonicator targeting 350 bp inserts. The resulting sheared DNA is selectively purified using sample purification beads to make the precise length of insert; End-repair (repaired to blunt end), A-tailing (A-base is added to 3'end), and ligation (Y-shaped adapter is used which includes a barcode) are performed as directed by TruSeq PCR-free Kit (Illumina, cat# FC-121-3003) protocols. A second Bead cleanup is performed after ligation to remove any residual reagents and adapter dimers. To verify the size of adapter-ligated fragments, we validate the template size distribution by running on a 2200 TapeStation (Agilent, Catalog # G2964AA) using a TapeStation DNA Screen Tape (Agilent, Catalog 5067-5588). The final libraries are quantified by qPCR assay using KAPA library quantification kit (cat.# KK4808 and KK4953) on a Light Cycler 480 instrument(Roche, cat# 05015278001).

Clustering and Sequencing

Eight normalized and indexed libraries were then pooled together and denatured before cluster generation on a cBot. The 8-plex pools were loaded on eight lanes of a flow cell and sequenced on a HiSeqX using illumina's HiSeq X ten reagents kit (V2.5, cat# FC-501-2521). For cluster generation, every step is controlled by cBot. When cluster generation is complete, the clustered patterned flow cells are then sequenced with sequencing software HCS (HiSeq Control Software). The runs are monitored for %Q30 bases using the SAV (Sequencing Analysis Viewer). Using RTA 2 (Real Time Analysis 2) the BCLs (base calls) were de-multiplexed into individual FASTQs per sample using illumina package bcl2fastq v2.15.0 and transferred from Macrogen to NWGC for alignment, merging, variant calling and sequencing QC.

Read Processing

Our processing pipeline consists of aligning FASTQ files to a human reference (hs37d5) using BWA-MEM (Burrows-Wheeler Aligner; v0.7.10) (Li and Durbin 2009). All aligned read data are subject to the following steps: (1) "duplicate removal" is performed, (i.e., the removal of reads with duplicate start positions; Picard MarkDuplicates; v1.111) (2) indel realignment is performed (GATK IndelRealigner; v3.2) resulting in improved base placement and lower false variant calls and (3) base qualities are recalibrated (GATK BaseRecalibrator; v3.2). Sample BAM files were "squeezed" using Bamutil with default parameters and checksummed before being transferred to the IRC.

Sequence Data QC

All sequence data undergo a QC protocol before they are released to the TOPMed IRC for further processing. For whole genomes, this includes an assessment of: (1) mean coverage; (2)

fraction of genome covered greater than 10x; (3) duplicate rate; (4) mean insert size; (5) contamination ratio; (6) mean Q20 base coverage; (7) Transition/Transversion ratio (Ti/Tv); (8) fingerprint concordance > 99%; and (9) sample homozygosity and heterozygosity. All QC metrics for both single-lane and merged data are reviewed by a sequence data analyst to identify data deviations from known or historical norms. Lanes/samples that fail QC are flagged in the system and can be re-queued for library prep (< 1% failure) or further sequencing (< 2% failure), depending upon the QC issue.

New York Genome Center

Soren Germer

DNA Sample Handling and QC

Genomic DNA samples were submitted in NYGC-provided 2D barcoded matrix rack tubes. Sample randomization was performed at investigator lab prior to sample submission. Upon receipt, the matrix racks were inspected for damage and scanned using a VolumeCheck instrument (BioMicroLab), and tube barcode and metadata from the sample manifest uploaded to NYGC LIMS. Genomic DNA was quantified using the Quant-iT PicoGreen dsDNA assay (Life Technologies) on a Spectramax fluorometer, and the integrity was ascertained on a Fragment Analyzer (Advanced Analytical). After sample quantification, a separate aliquot (100ng) was removed for SNP array genotyping with the HumanCoreExome-24 array (Illumina). Array genotypes were used to estimate sample contamination (using VerifyIDintensity), for sample fingerprinting, and for downstream quality control of sequencing data. Investigator was notified of samples that failed QC for total mass, degradation or contamination, and replacement samples were submitted.

Library Construction

Sequencing libraries were prepared using the TruSeq PCR-free DNA HT Library Preparation Kit (Illumina) with 500 ng DNA input, following manufacturer's protocol with minor modifications to account for automation. Briefly, genomic DNA was sheared using the Covaris LE220 sonicator to a target size of 450 bp (t:78; Duty:15; PIP:450; 200 cycles), followed by end-repair and bead based size selection of fragmented molecules. The selected fragments were A-tailed, and sequence adaptors ligated onto the fragments, followed by two bead clean-ups of the libraries. These steps were carried out on the Caliper SciClone NGSx workstation (Perkin Elmer). Final libraries are evaluated for size distribution on the Fragment Analyzer and quantified by qPCR with adaptor specific primers (Kapa Biosystems).

Clustering and Sequencing

Final libraries were multiplexed for 8 samples per sequencing lane, with each sample pool sequenced across 8 flow cell lanes. 1% PhiX control was spiked into each library pool. The library pools were quantified by qPCR, loaded on the to HiSeq X patterned flow cells and clustered on an Illumina cBot following manufacturer's protocol. Flow cells were sequenced on the Illumina HiSeq X with 2x150bp reads, using V2 sequencing chemistry, and Illumina HiSeq Control Software v3.1.26.

Read Processing

Demultiplexing of sequencing data was performed with bcl2fastq2 v2.16.0.10, and sequencing data was aligned to human reference build 37 (hs37d5 with decoy) using BWA-MEM v0.7.8. Data was further processed using the GATK best-practices v3.2-2 pipeline, with duplicate marking using Picard tools v1.83, realignment around indels, and base quality recalibration. Individual sample BAM files were squeezed using Bamutil v1.0.9 with default parameters -- removing OQ's, retaining duplicate marking and binning quality scores (binMid) -- and transferred to the IRC using Globus. Individual sample SNV and indel calls were generated using GATK haplotype caller and joint genotyping was performed across all the NYGC phase 1 samples.

Sequence Data QC

Prior to release of BAM files to IRC, we ensured that mean genome coverage was >=30x, when aligning to the ~2.86Gb sex specific mappable genome, and that uniformity of coverage was acceptable (>90% of genome covered >20x). Sample identity and sequencing data quality was confirmed by concordance to SNP array genotypes. Sample contamination was estimated with VerifyBAMId v1.1.0 (threshold <3%). Gender was determined from X- and Y-chromosome coverage and checked against submitter information. Further QC included review of alignment rates, duplicate rates, and insert size distribution. Metrics used for review of SNV and indel calls included: the total number of variants called, the ratio of novel to known variants, and the Transition to Transversion ratios, and the ratio of heterozygous to homozygous variant calls.

Illumina Genomic Services

Karine Viaud Martinez

DNA Sample Handling and QC

Project samples are processed from 96-well barcoded plates provided by Illumina. Electronic manifest including unique DNA identification number describing the plate barcode and well position (eg, LP6002511-DNA_A01) and samples information (e.g. Gender, Concentration, Volume, Tumor/normal, Tissue type, Replicate…) is accessioned in LIMS. This enables a seamless interface with our robotic processes and retains sample anonymity. An aliquot of each sample is processed in parallel through the Infinium Omni 2.5M (InfiniumOmni2.5Exome-8v1, HumanOmni25M-8v1) genotyping array and an identity check is performed between the sequencing and array data via an internal pipeline. Genomic DNA is quantified prior to library construction using PicoGreen (Quant-iT™ PicoGreen® dsDNA Reagent, Invitrogen, Catalog #: P11496). Quants are read with Spectromax Gemini XPS (Molecular Devices).

Library Construction

Samples are batched using LIMS, and liquid handling robots perform library preparation to guarantee accuracy and enable scalability. All sample and reagent barcodes are verified and recorded in LIMS. Paired-end libraries are generated from 500ng–1ug of gDNA using the

Illumina TruSeq DNA Sample Preparation Kit (Catalog #: FC-121-2001), based on the protocol in the TruSeq DNA PCR-Free Sample Preparation Guide. Pre-fragmentation gDNA cleanup is performed using paramagnetic sample purification beads (Agencourt® AMPure® XP reagents, Beckman Coulter). Samples are fragmented and libraries are size selected following fragmentation and end-repair using paramagnetic sample purification beads, targeting short insert sizes. Final libraries are quality controlled for size using a gel electrophoretic separation system and are quantified.

Clustering and Sequencing

Following library quantitation, DNA libraries are denatured, diluted, and clustered onto v4 flow cells using the Illumina cBot™ system. A phiX control library is added at approximately 1% of total loading content to facilitate monitoring of run quality. cBot runs are performed based on the cBot User Guide, using the reagents provided in Illumina TruSeq Cluster Kit v4.
Clustered v4 flow cells are loaded onto HiSeq 2000 instruments and sequenced on 125 bp paired-end, non-indexed runs. All samples are sequenced on independent lanes. Sequencing runs are performed based on the HiSeq 2000 User Guide, using Illumina TruSeq SBS v4 Reagents. Illumina HiSeq Control Software (HCS) and Real-Time Analysis (RTA) are used on HiSeq 2000 sequencing runs for real-time image analysis and base calling.

Read Processing

The Whole Genome Sequencing Service leverages a suite of proven algorithms to detect genomic variants comprehensively and accurately. Most versions of the Illumina callers are open source and available publicly. See the Illumina GitHub (https://github.com/Illumina) for the current releases. One of more lanes of data were processed from run folders directly with the internal use only ISAS framework (2.5.55.16 or 2.5.26.13 depending on the start of the project), including alignment with iSAAC (iSAAC-01.14.02.06 or iSAAC-SAAC00776.15.01.27), small variant called with Starling (2.0.17 or starka-2.1.4.2), structural variant called with Manta (manta-0.18.1 or manta-0.23.1) and copy number variant Canvas (v4.0).

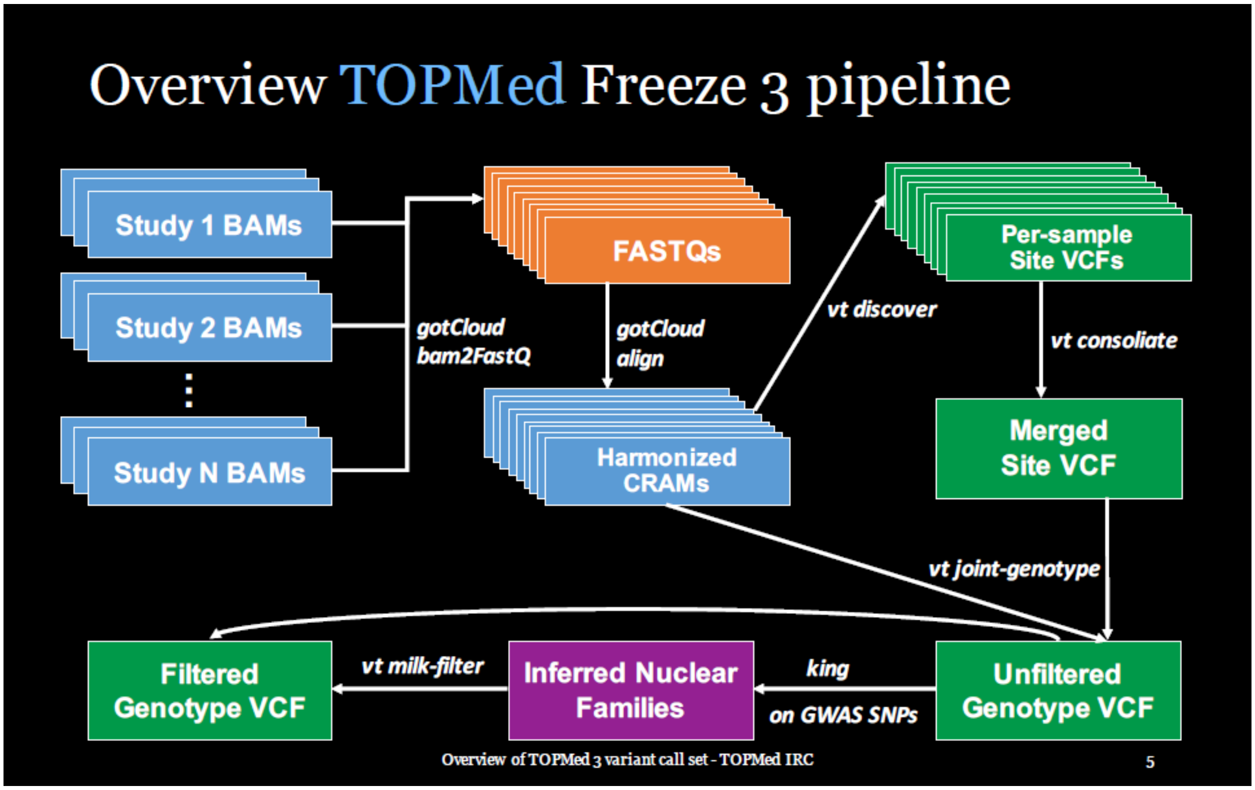
Sequence Data QC

The genome build QC pipeline is automated to evaluate both primary (sequencing level) and secondary (build level) metrics against expectations based on historical performance. Multiple variables, such as Gb of high quality (Q30) data, mismatch rates, percentage of aligned reads, insert size distribution, concordance to the genotyping array run in parallel, average depth of coverage, number of variants called, callability of the genome as a whole as well as of specific regions (evenness of coverage), het/hom ratio, duplicate rates, and noise are assessed. Genome builds that are flagged as outliers at QC are reviewed by our scientists for investigation. Scientists review all QC steps during the process: Library quantification and fragment size; run quality; genotyping and sequencing data considering Sample Manifest information (Tumor/Normal, tissue type). Libraries or sequencing lanes may be requeued for additional sequencing or library prep as needed.

Informatics Research Center Methods

Tom Blackwell, Hyun Min Kang and Goncalo Abecasis
Center for Statistical Genetics, Department of Biostatistics, University of Michigan

The IRC pipeline consists of two major processes diagrammed in the Figure below: (1) Harmonization of data from the BAM files provided by the Sequencing Centers and (2) joint variant discovery and genotype calling across studies. Detailed protocols for these processes are given in the following sections.



Overview TOPMed Freeze 3 pipeline

Study 1 BAMs
Study 2 BAMs
Study N BAMs
gotCloud bam2FastQ
FASTQs
gotCloud align
Harmonized CRAMs
vt discover
Per-sample Site VCFs
vt consoliate
Merged Site VCF
vt joint-genotype
Unfiltered Genotype VCF
king on GWAS SNPs
Inferred Nuclear Families
vt milk-filter
Filtered Genotype VCF

Overview of TOPMed 3 variant call set - TOPMed IRC 5

Harmonization of Read Alignments

Prior to joint variant discovery and genotype calling by the IRC, the sequence data obtained from the TOPMed Sequencing Centers are remapped using a standard protocol to produce “harmonized” BAM files.

Sequence data are received from each sequencing center in the form of .bam files mapped to the 1000 Genomes hs37d5 build 37 decoy reference sequence. File transfer is via Aspera or Globus Connect, depending on the center. Batches of 100 - 500 .bam files in a single directory are convenient, along with a file of md5 checksums for the files in that directory.

The IRC validates the md5 checksum, indexes each .bam file using ‘samtools index’ and runs local programs Qplot (Li, et al, 2013, doi:10.1155/2013/865181) and verifyBamId (Jun, et al, 2012, doi:10.1016/j.ajhg.2012.09.004) for incoming sequence quality control. We make a backup copy in .cram format using ‘samtools view -C’ with base call quality scores reduced to 8 bins using ‘bamUtils squeeze’ (if not already binned at the sequencing center). We add ‘’NWD’’ DNA sample identifiers to the read group header lines (Illumina) and convert from UCSC to Ensembl chromosome names (Illumina and Macrogen) using ‘samtools reheader’. In house scripts add read group tags as needed to legacy Illumina sequencing data from 2012-2013.

To produce the ‘’harmonized’’ read mappings which are deposited in dbGaP and used for variant discovery and genotyping, we remap the sequence data in each .bam file to the 1000 Genomes hs37d5 decoy reference sequence using a uniform ‘’IRC standard’’ protocol. This uses ‘bamUtils bam2fastq’ with flags ‘--splitRG --gzip’ to extract all sequence reads by read group into paired-end .fastq format, then remaps to hs37d5.fa using ‘bwa mem’ version 0.7.12-r1039 with ‘-M’ to mark split alignments as secondary. Read group header information is copied from the sequencing center .bam file. Followed by ‘samtools sort’, ‘bamUtils polishBam’, ‘bamUtils mergeBam’ and ‘bamUtils dedup_LowMem --recab’ with flags ‘--binMid --binQualS 2,3,10,20,25,30,35,40,50 --maxBaseQual 44’ to recalibrate and bin base call quality scores. Samtools version 1.2 is used throughout. Processing is coordinated and managed by our ‘GotCloud’ processing pipeline.

Description of our local and standard software tools is available from:

http://genome.sph.umich.edu/wiki/BamUtils
http://genome.sph.umich.edu/wiki/GotCloud
http://genome.sph.umich.edu/wiki/QPLOT
http://www.htslib.org (samtools)
https://github.com/lh3/bwa (bwa, current)
http://bio-bwa.sourceforge.net (bwa, outdated)

Software sources:

http://genome.sph.umich.edu/w/images/7/70/BamUtilLibStatGen.1.0.13.tgz
https://github.com/statgen/bamUtil/releases/tags/v1.0.13
https://github.com/statgen/gotcloud/releases/tags/gotcloud.1.17.4
http://www.sph.umich.edu/csg/zhanxw/software/qplot/qplot-source.20130627.tar.gz
https://github.com/statgen/verifyBamId/releases/tags/v1.1.2
https://github.com/samtools/samtools/releases/download/1.2/samtools-1.2.tar.bz2
https://sourceforge.net/projects/bio-bwa/files/bwakit/bwakit-0.7.12_x64-linux.tar.bz2
https://github.com/lh3/bwa (source code)

GRCh37 genome reference source:

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/hs37d5.fa.gz

(Note that the original 1000 Genomes .fasta file is razip compressed. For use with the current version of samtools, this file must be expanded and (optionally) re-compressed with bgzip to support the current samtools faidx indexing.)

The two sequence quality criteria we use in order to pass sequence data on for joint variant discovery and genotyping are: estimated DNA sample contamination below 3%, and fraction of the genome covered at least 10x 95% or above. DNA sample contamination is estimated from the sequencing center read mapping using software verifyBamId (Goo Jun, et al. (2012) Detecting and estimating contamination of human DNA samples in sequencing and array based genotype data. American Journal of Human Genetics, v.91, n.5, pp.839-848).

The IRC-harmonized BAM files and the original BAM files from the Sequencing Centers are deposited in the NCBI Sequence Read Archive, where they are stored in '.sra' format. These files can be accessed by approved users through the dbGaP "Run Selector". Note that the two different read mappings (IRC-harmonized versus Sequencing Center) can be distinguished in the Run Selector by the column ''Alignment Provider''. Users download .sra files using the SRA ToolKit. Each sample is identified to the toolkit with an ID starting with "SRR", which can be matched with the sample ID (NWD ID) using the dbGaP Run Selector. Users can request all data for a sample, or just a specific region of the genome, and easily convert from SAM back to BAM format using samtools. The SRA ToolKit documentation includes an example of how to create a BAM file for a specified genomic region for a given sample.

SRA ToolKit setup: https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=std
SRA ToolKit .sra to .bam for specified region:
https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=sam-dump
samtools: http://www.htslib.org/

Variant Discovery and Genotype Calling

Overview

The genotype call sets provided for the dbGaP accessions released in Oct 2016 are from "freeze 3a" of the variant calling pipeline performed by the TOPMed Informatics Research Center (Center for Statistical Genetics, University of Michigan, Hyun Min Kang, Tom Blackwell and Goncalo Abecasis). The software tools used in this version of the pipeline are available in the following repository: https://github.com/statgen/topmed_freeze3_calling. The following description refers to specific components of the pipeline. The variant calling software tools are under active development; updated versions can be accessed at http://github.com/atks/vt or http://github.com/hyunminkang/apigenome.

Outline of the variant calling procedure

The `GotCloud vt` pipeline detects variant sites and calls genotypes from a list of aligned sequence reads. Specifically, the pipeline consists of the following six key steps (see also the Figure). Most of these procedure will be integrated into GotCloud software package later in 2016.

1. **Variant detection** : For each sequenced genome (in BAM/CRAMs), candidate variants are detected by `vt discover2` software tools, separated by each chromosome. The candidate variants are normalized by `vt normalize` algorithm.
2. **Variant consolidation** : For each chromosome, the called variant sites are merged across the genomes, accounting for overlap of variants between genomes, using `vt consolidate` software tool.
3. **Genotype and feature collection** : For each 100kb chunk of genome, the genotyping module implemented in `vt joint_genotype_sequential` collects individual genotypes and variant features across the merged sites by iterating over sequenced genomes, focusing on the selected region.
4. **Variant filtering** : We use the inferred pedigree of related and duplicated samples to calculate Mendlian consistency statistics using `vt milk-filter`, and to train a variant quality classifier using a Support Vector Machine (SVM) implemented in the `libsvm` software package.

Steps to prepare input files, install software and perform variant calling

To produce variant calls using this pipeline, the following input files need to be prepared:

1. Aligned sequenced reads in BAM or CRAM format. Each BAM and CRAM file should contain one sample per subject. It also must be indexed using `samtools` index or equivalent software tools.
2. A sequence index file. Each line should contain [Sample ID] [Full Path to the BAM/CRAM file] [Contamination Estimates -- put zero if unknown]. See `data/trio_data.index` for example.
3. A pedigree file of nuclear families and duplicates in PED format. The pedgiree file should contain only nuclear families. When a sample is duplicated, all Sample IDs representing the same individual (in the 2nd column) need to presented in a comma-separated way. In the 3rd and 4th column to represent their parents, only a representative sample ID is required. See `data/trio_data.ped` for example.

To clone and build the repository, follow these steps

```
  $ git clone https://github.com/statgen/topmed_freeze3_calling.git
  $ cd topmed_freeze3_calling
  $ make  # or make -j [numjobs] to expedite the process
  $ wget ftp://anonymous@share.sph.umich.edu/gotcloud/ref/hs37d5-
db142-v1.tgz  # this will take a while
  $ tar xzvf hs37d5-db142-v1.tgz
  $ rm hs37d5-db142-v1.tgz
```

After these steps, modify scripts/gcconfig.pm to specify input data files or other parameters. Modifying the first section (index and ped file in particular) should be minimally required changes.

To perform variant discovery and consolidation, run the following step
 `$ perl scripts/step1-detect-and-merge-variants.pl` [whitespace separated chromosome names to call]
After this step, follow the instruction to run `make -f [Makefile] -j [numjobs]` to complete the discovery tasks

To genotype variants, run the following steps.
 `$perl scripts/step2-joint-genotyping.pl` [whitespace separated chromosome names to call]
After this step, following the instruction to `run make -f [Makefile] -j [numjobs]` to complete the discovery tasks

To perform variant filtering using pedigree information, follow these steps.
  `$ perl scripts/step3a-compute-milk-score.pl` [whitespace separated chromosome names to call] ## run makefile after this step
 `$ perl scripts/step3b-run-svm-milk-filter.pl` [whitespace separated chromosome names to call]
 `$ perl scripts/step3c-run-milk-transfer.pl` [whitespace separated chromosome names to call] ## this step is needed only when performing transfer learning from other chromosomes.
After all these steps, the called variant sites will be available at `$(OUTPUT_DIR)/svm`, and the genotypes will be available at `$(OUTPUT_DIR)/paste`.

Variant Detection

Variant detection from each sequenced (and aligned) genome is performed by `vt discover2` software tool. The script `step-1-detect-variants.pl` provides a means to automate the variant detection across a large number of sequenced genomes.
The variant detection algorithm considers a variant as a potential candidate variant if there exists a mismatch between the aligned sequence reads and the reference genome. Because

such a mismatch can easily occur by random errors, only potential candidate variants passing the following criteria are considered to be ***candidate variants*** in the next steps.

1. At least two identical evidences of variants must be observed from aligned sequence reads.
    a. Each individual evidence will be normalized using the normalization algorithm implemented in `vt normalize` software tools.
    b. Only evidence from the reads with mapping quality 20 or greater will be considered.
    c. Duplicate reads, QC-failed reads, supplementary reads, secondary reads will be ignored.
    d. Evidence of a variant within overlapping fragments of read pairs will not be double counted. Either end of the overlapping read pair will be soft-clipped using bam `clipOverlap` software tool.
2. Assuming per-sample heterozygosity of 0.1%, the posterior probability of having a variant at the position should be greater than 50%. This method is equivalent to the `glfSingle` model described in http://www.ncbi.nlm.nih.gov/pubmed/25884587

The variant detection step is required only once per sequenced genome, when multiple freezes of variant calls are produced over the course of time.

Variant Consolidation

Variants detected from the discovery step are merged across all samples. This step is implemented in the `step-2-detect-variants.pl` scripts.

1. The non-reference alleles normalized by `vt normalize` algorithm are merged across the samples, and unique alleles are printed as biallelic candidate variants. The algorithm is published at http://www.ncbi.nlm.nih.gov/pubmed/25701572
2. If there are alleles overlapping with other SNPs and Indels, `overlap_snp` and `overlap_indel` filters are added in the FILTER column of the corresponding variant.
3. If there are tandem repeats with 2 or more repeats with total repeat length of 6bp or longer, the variant is annotated as a potential VNTR (Variant Number Tandem Repeat), and `overlap_vntr` filters are added to the variant overlapping with the repeat track of the putative VNTR.

Variant Genotyping and Feature Collection

The genotyping step iterates all of the merged variant site over the sequenced samples. It iterates over BAM/CRAM files one at a time sequentially for each 1Mb chunk to perform contamination-adjusted genotyping and annotation of variant features for filtering. The following variant features are calculated during the genotyping procedure.

- AVGDP : Average read depth per sample
- AC : Non-reference allele count
- AN : Total number of alleles

- GC : Genotype count
- GN : Total genotype counts
- HWEAF : Allele frequency estimated from PL under HWE
- HWDAF : Genotype frequency estimated from PL under HWD
- IBC : [Obs(Het) – Exp(Het)] / Exp[Het]
- HWE_SLP : -log(HWE likelihood-ratio test p-value) ✕ sign(IBC)
- ABE : Average fraction [#Ref Allele] across all heterozygotes
- ABZ : Z-score for tesing deviation of ABE from expected value (0.5)
- BQZ: Z-score testing association between allele and base qualities
- CYZ: Z-score testing association between allele and the sequencing cycle
- STZ : Z-score testing association between allele and strand
- NMZ : Z-score testing association between allele and per-read mismatches
- IOR : log [Obs(non-ref, non-alt alleles) / Exp(non-ref, non-alt alleles)]
- NM1 : Average per-read mismatches for non-reference alleles
- NM0 : Average per-read mismatches for reference alleles

The genotyping is done with adjustment for potential contamination. It uses adjusted genotype likelihood similar to the published method https://github.com/hyunminkang/cleancall, but does not use estimated population allele frequency for the sake of computational efficiency. It conservatively assumes that the probability of observing a non-reference read given a homozygous reference genotype is equal to half of the estimated contamination level, (or 1%, whichever is greater). The probability of observing a reference read given a homozygous non-reference genotype is calculated in a similar way. This adjustment makes the heterozygous call more conservatively when the reference and non-reference allele reads are strongly imbalanced. For example, if 45 reference alleles and 5 non-reference alleles are observed at Q40, the new method calls it as homozygous reference genotype while the original method ignoring potential contamination calls it as heterozygous genotype. This adjustment improves the genotype quality of contaminated samples by reducing genotype errors by several fold.

Variant Filtering

The variant filtering in TOPMed Freeze 3 were performed by (1) first calculating Mendelian consistency scores using known familial relatedness and duplicates, and (2) training SVM classifier between the known variant sites (positive labels) and the Mendelian inconsistent variants (negative labels).

The negative labels are defined if the Bayes Factor for Mendelian consistency quantified as `Pr(Reads | HWE, Pedigree) / Pr(Reads | HWD, no Pedigree)` is less than 0.001. Also a variant is marked as negative labels if 3 or more samples show 20% of non-reference Mendelian discordance within families or genotype discordance between duplicated samples. The positive labels are the SNPs found polymorphic either in the 1000 Genomes Omni2.5 array or in HapMap 3.3, with additional evidence of being polymorphic from the sequenced samples. Variants eligible to be marked with both positive and negative labels are discarded from the labels. The SVM scores trained and predicted by the libSVM software tool are annotated in the VCF file.

Two additional hard filters were applied. (1) Excess heterozygosity filter (EXHET), if the Hardy-Weinberg disequilbrium p-value was less than 1e-6 in the direction of excess heterozygosity. An additional ~3,900 variants were filtered out by this filter. (2) Mendelian discordance filter (DISC), with 3 or more Mendelian inconsistencies or duplicate discordances observed from the samples. An additional ~370,000 variants were filtered out by this filter.

Functional annotation for each variant is provided in the INFO field using Pablo Cingolani's snpEff 4.1 with a GRCh37.75 database. The current release includes only hard-call genotypes in the VCF files, without genotype likelihoods and with no missing genotypes. Genotype likelihoods may be included in future releases, at the cost of approximately 100x greater file size.

Data Coordinating Center Methods

Cathy Laurie, Bruce Weir and Ken Rice
Genetic Analysis Center, Department of Biostatistics, University of Washington

The following three approaches were used to identify and resolve sample identity issues.

Concordance between annotated sex and biological sex inferred from the WGS data

Biological sex was inferred from normalized X and Y chromosome depth for each sample (i.e. divided by autosomal depth) and from X chromosome heterozygosity. A small number of sex mismatches were detected as annotated females with low X and high Y chromosome depth or annotated males with high X and low Y chromosome depth. These samples were either excluded from the sample set to be released on dbGaP or their sample identities were resolved using information from array comparisons or pedigree checks. We also identified a small number of sex chromosome aneuploidies (XXY, XXX and mosaics such as XX/XO); these samples are excluded from the October 2016 release, but will be annotated and included in subsequent releases.

Concordance between prior SNP array genotypes and WGS-derived genotypes

Prior genome-wide SNP array data are available for 12 of the 16 accessions to be released in Oct 2016 (all except phs001024, phs001062, phs001032, and phs000997). The average percentage of individuals within those 12 accessions who have prior array data is 97%.

For five accessions, the prior array data analyzed for TOPMed were derived from 'fingerprints' compiled by dbGaP (Yumi Jin, see URL below). The fingerprints consist of genotypes from a set of 10,000 bi-allelic autosomal SNP markers chosen to occur on multiple commercial arrays and to have a minor allele frequency (MAF) > 5%. For the remaining seven accessions, all autosomal SNPs with MAF > 5% on a genome-wide array were used. In both cases (fingerprints and full array), percent concordance was determined by matching on heterozygous versus homozygous status (rather than specific alleles) to avoid strand issues. Concordance percentages for array-WGS matches were generally in the high 90s, while those considered to be mismatches were in the 50-60% range (empirically determined to be the expected matching level for random pairs of samples). We found that 99.6% of the 12,386 WGS samples tested were concordant with prior array data. Discordant samples were either excluded from the October 2016 release or resolved as sample switches using pedigree and/or sex-mismatch results.

SNP fingerprints: http://www.ashg.org/2014meeting/abstracts/fulltext/f140122979.htm

Comparisons of observed and expected relatedness from pedigrees

Kinship coefficients (KCs) were estimated for all pairs of individuals using ~250k single nucleotide variants that are autosomal, MAF >5%, heterozygosity < 0.55, and pruned to have low linkage disequilibrium (r^2<0.1) with one another. The estimation procedure used 'PC-Relate' (Conomos et al. 2016, DOI: 10.1016/j.ajhg.2015.11.022), which is robust to population structure, admixture and departures from Hardy-Weinberg Disequilibrium. The KC estimates were compared to those expected from pedigrees for the accessions with annotated family structure (phs000956, phs000974, phs000988, phs000964, and phs000954). Discrepancies between observed and expected KCs were investigated and, in many cases, resolved either by correcting sample-subject mapping for sample switches or by making a change in the pedigree structure. Pedigree changes were warranted when one alteration resolved multiple KC discrepancies or when supported by additional information from the studies.